Advanced Text Mining Methods

Carsten Jentsch, Kai-Robin Lange

TU Dortmund, Department of Statistics

01/24/2024

technische universität
dortmund

fakultät
statistik

Schedule and Requirements

Requirements

- Presentation in English; 30+10/45+10 minutes
- Report up to 12 pages long in English or German
- Open for Bachelor & Master Statistics, Data Science, Econometrics
- Basic knowledge of Natural Language Processing is expected.

Schedule

- Distribution of topics: voting for priorities until 04/01/2024
- Presentations: August 2024 (tbd)
- Reports due: 09/09/2024

Motivation

Research Questions

- What do people talk about? – Topic Models
- How do words relate to each other? – Word embeddings
- How can we train a large neural network with little data? – Few-shot learning
- Developement of a word's meaning over time? – Diachronic embeddings
- Changing texts over time? – Change point detection

Projects

Goal: deeper knowledge on modern developements in NLP research

Theory-oriented works

- Diachronic embeddings
- Pattern-exploiting training (PET) and its modern developements
- Few-shot learning techniques for Transformers
- Open Source LLMs
- Neural Topic Models
- Dynamic Topic Models
- ...

Practice-oriented pipelines

- Diachronic Embeddings
- Change point detection
- ...